*Studies in Second Language Acquisition* (2024), **46**, 1069–1093
doi:10.1017/S0272263124000408

RESEARCH ARTICLE

# Differential effects of identification and discrimination training tasks on L2 vowel identification and discrimination

Juli Cebrian[1], Núria Gavaldà[2], Celia Gorba[1] and Angélica Carlet[3]

[1]Universitat Autònoma de Barcelona; [2]Universitat de Barcelona and [3]Charles Darwin University
**Corresponding author:** Juli Cebrian; Email: juli.cebrian@uab.cat

(Received 01 June 2023; Revised 04 May 2024; Accepted 20 May 2024)

## Abstract

High variability phonetic training using perceptual tasks such as identification and discrimination tasks has often been reported to improve L2 perception. However, studies comparing the efficacy of different tasks on different measures are rare. Forty-four Catalan/Spanish bilingual learners of English were trained with identification or categorical discrimination tasks and were tested on both measures. Results showed that both methods were successful in improving the identification and discrimination of English vowels. Training with nonword stimuli generalized to new nonwords and real word stimuli, and improvement was maintained four months later. Cross-task effects may be related to the categorical nature of the discrimination task, which may entail a level of processing similar to that of identification training. Interestingly, whereas identification training improved identification more than discrimination training, discrimination training did not enhance discrimination more than identification training. This asymmetry may be explained by task differences in the amount and type of feedback used.

**Keywords:** high variability phonetic training (HVPT); identification; discrimination; generalization; retention

## Introduction

Acquiring novel L2 sounds in a foreign language context can be challenging particularly given the probable scarcity of authentic target language input available. Against this background, a possible source of specialized target language experience can be found in high variability phonetic training (HVPT), which exposes learners to highly variable stimuli (i.e., a variety of talkers, stimuli, and phonetic contexts) to provide them with the kind of variability present in real communicative situations. Thus, HVPT allows learners to attend to the aspects of the stimuli that are crucial to identifying and

distinguishing L2 categories and to disregard talker-specific or context-specific characteristics (Lively, Logan & Pisoni, 1993). In addition, HVPT is used to draw learners' attention to particularly challenging target structures through the use of immediate corrective feedback (Logan & Pruitt, 1995; Thomson, 2018). The efficacy of phonetic training is typically assessed by contrasting the trained learners' performance before and after training and also in comparison with a control group of untrained learners. Furthermore, HVPT is expected to promote the generalization of learning to untrained structures, such as new talkers, sounds, stimuli, and phonetic contexts, which is believed to indicate the formation of robust L2 categories (Logan & Pruitt, 1995). The results of numerous studies generally support the efficacy of HVPT for enhancing the perception and production of L2 sounds, and promoting generalization of knowledge (see Thomson, 2018, for an overview). For instance, several studies comparing HVPT with low variability phonetic training (LVPT, stimuli from a single talker) have shown that only the former results in generalization of learning (Lively et al., 1993; Perrachione, Lee, Ha & Wong, 2011), although some recent studies have questioned this advantage of HVPT over LVPT and point to other factors in addition to talker variability that may contribute to generalization (Brekelmans, Lavan, Saito, Clayards & Wonnacott, 2022; Zhang, Cheng & Zhang 2021a; Zhang, Cheng, Qin & Zhang 2021b). Another measure of robust learning is the degree to which the improvement obtained through training is maintained for some time after training has ended. This is referred to as retention and is typically measured by means of a delayed test, equal to the pretests and posttests. HVPT studies have found evidence of retention of learning up to three months after training (Lively, Pisoni, Yamada, Tohkura & Yamada, 1994), and very few report retention after a longer period (Iverson & Evans, 2009; Thomson, 2018). The current study thus explores the potential of HVPT further by contrasting the effect of two different perceptual training methods, namely identification and discrimination tasks, on both the ability to identify and to discriminate target language sounds, within a single study, unlike most previous works. In addition, the effectiveness of the training tasks is evaluated by analyzing if the training methods result in generalization and retention of learning. Hence, the learners' ability to identify and discriminate target sounds in stimuli not used in training (new nonword and real word stimuli produced by new talkers) is examined right before and after training, and four months after training. The characteristics of the two training methods are described next.

### *Perceptual training tasks*

Most perceptual training studies make use of identification (ID) tasks (e.g., Lengeris & Hazan, 2010; Iverson, Pinet & Evans, 2012), which require listeners to identify or label a given aural stimulus; some use discrimination (DIS) tasks (e.g., Strange & Dittman, 1984; Georgiou, 2021), in which listeners indicate if two (or more) aural stimuli belong to the same category or not, and some have used a combination of perceptual tasks (e.g., Shinohara & Iverson, 2018, 2021). Yet, few studies have actually compared the effectiveness of different tasks on different abilities in the same study. Generally, identification (ID) tasks have been found to improve identification (e.g., Lambacher, Martens, Kakehi, Marasinghe & Molholt, 2005; Iverson & Evans, 2009) and discrimination (DIS) tasks successfully improve discrimination (Georgiou, 2021). Some studies examining cross-task effects have reported a greater efficacy of ID training. For instance, Jamieson and Morosan (1986) found that identification training resulted in improved identification and discrimination of trained synthetic stimuli as well as of

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

untrained natural stimuli. By contrast, Strange and Dittman (1984) found that DIS training improved the identification and the discrimination of the English /r/-/l/ contrast but did not result in the generalization of learning to natural stimuli. However, these early studies used synthetic stimuli and did not involve high variability, which may account for the limited results reported for DIS training. Some later studies that employed HVPT have reported that ID training successfully improves identification but has little effect on discrimination (Lengeris & Hazan, 2010; Iverson et al., 2012). Lengeris and Hazan (2010) found that identification training with natural stimuli improved the identification but not the discrimination of L2 English vowels by Greek speakers. The fact that identification was tested under different conditions (natural and synthetic stimuli) but discrimination was tested only using synthetic stimuli may explain the lack of an effect of ID training on discrimination. Regarding Iverson et al. (2012), the advantage of ID training over DIS training may be the result of procedural learning for the former, as pre- and posttraining tests examined identification only, giving a task familiarity advantage to ID trainees.

Arguments for the apparent advantage of ID have also been linked to a difference in the nature of the two tasks, as DIS may draw listeners' attention to variability within the same category and tap into lower levels of phonological processing, while ID focuses on variability between categories and involves higher levels of phonological encoding that may be more relevant for L2 categorization (Jamieson & Morosan, 1986; Logan & Pruitt, 1995; Iverson et al., 2012). This difference is particularly notable when discrimination tasks involve auditory discrimination, that is, when the same trials contain physically identical stimuli, and different trials may involve physically different stimuli from the same phoneme category (Polka, 1992; Strange, 1992). Discrimination tasks that are categorical in nature, that is, where the same trials consist of physically different stimuli representative of the same category (e.g., different productions by the same speaker or by different speakers), may in fact involve a similar level of processing as identification tasks. To explore this further, the current study compares the use of ID tasks and specifically categorical DIS tasks in HVPT.

Few studies have directly compared the effect of ID and DIS training in the same study. Flege (1995) examined whether a categorical AX discrimination task or a two-alternative forced choice identification task was more effective for training Mandarin learners of English to identify English word-final /d/ and /t/. The results indicated that both types of tasks were equally effective and led to the generalization of learning to untrained stimuli, contrary to earlier findings using auditory discrimination tasks (e.g., Strange & Dittman, 1984). Similarly, other studies have provided evidence for the effectiveness of both ID and categorical DIS tasks in improving the discrimination of Thai tones (Wayland & Li, 2008), the use of cue-weighting in the perception of the English /iː/-/ɪ/ contrast (Wee, Grenon, Sheppard & Archibald, 2019), and the identification and discrimination of the English /r/ and /l/ contrast (Shinohara & Iverson, 2018).

However, divergent results for ID training and categorical DIS training have also been reported. In a study involving Japanese learners of English, Nozawa (2015) compared the effect of ID training and categorical AXB DIS training on the identification of English vowels and coda nasals. In this case, the results showed that the tasks had a comparable positive effect of both ID and DIS training in the case of coda nasals, but ID training yielded better results with vowel identification. A similar finding was reported by Carlet and Cebrian (2022), who also found a greater benefit of ID training for vowel identification but comparable effects of ID and DIS training on stop identification. Greater improvement with ID tasks has also been reported for the

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

perception of the /z/ vs. /dz/ contrast in coda position (i.e., *rose* vs. *roads*, Law, Grenon, Sheppard & Archibald, 2019). On the other hand, Carlet and Cebrian (2022) also report an example of a possible benefit of DIS tasks. Their study explored the effect of training on implicitly exposed but untargeted sounds, in addition to the specifically targeted sounds; two groups were trained on vowels and two other groups were trained on stops, with the same set of CVC stimuli, and all groups were tested on all (i.e., targeted and untargeted) sounds. Interestingly, only the AX DIS training led to an enhanced perception of untargeted L2 sounds. The authors provide several explanations for this difference, including the possibility that, unlike ID training, which directs the listeners' attention to the sound that is to be identified, DIS training may allow listeners to attend to the whole stimulus, paying attention to other sounds present in the stimulus in addition to the targeted sounds.

In brief, previous studies comparing ID and (categorical) DIS tasks show comparable results (Flege, 1995; Shinohara & Iverson, 2018) or a certain advantage of ID training (Nozawa, 2015; Carlet & Cebrian, 2022). Yet, comparisons across studies are complicated due to the differences in study design. For instance, neither Nozawa (2015) nor Wee et al. (2019) included a control group or a test of generalization or retention. Further, except for Shinohara and Iverson (2018), who tested both identification and discrimination, and Wayland and Li (2008), who tested discrimination only, most studies used only identification tasks at the pretest and posttest (Flege, 1995; Nozawa, 2015; Carlet & Cebrian, 2022), which may also have contributed to the advantage of ID training due to procedural learning. Finally, while Shinohara and Iverson (2018) compared ID and DIS training on both identification and discrimination abilities, the study did not include a control group and discrimination training included both auditory discrimination and categorical discrimination tasks. The current study thus contrasts the effect of ID and DIS (specifically, categorical DIS) on both the ability to identify and discriminate L2 vowels, including a control group, and assessing generalization and retention of learning.

Generalization is examined in the current study by evaluating the learners' ability to identify and discriminate L2 vowels in new nonwords and real words produced by new talkers after undergoing training with nonword stimuli. The use of nonword training stimuli responds to the need to avoid the potential effects of word familiarity and orthographic interference found with real words. In fact, previous works show that phonetically-oriented training using nonwords (as opposed to lexically-oriented training with real words) may be more efficient at forcing the trainees' attention to the important phonetic details that facilitate the perception of different L2 categories, thus improving L2 perception (Carlet & Cebrian, 2022) and production (Thomson & Derwing, 2016; Ortega, Mora-Plaza & Mora, 2021; Mora, Ortega, Mora-Plaza & Aliaga-García, 2022). On the other hand, studies also indicate that perception of L2 contrasts may be facilitated when sounds are presented in a lexical context. For instance, previous studies reported that adult L2 learners were better at discriminating (Mora, 2005) and identifying (Rato & Carlet, 2020) challenging L2 phones in real words than in nonwords, showing that lexical representations may play a role in the perception of segmental L2 contrasts (Yamada, Tohkura & Kobayashi, 1997).

## *The present study*

The main purpose of the current study is to examine the effectiveness of two perceptual training tasks (identification [ID] and categorical discrimination [DIS]) for training

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

Spanish/Catalan-speaking learners of L2 English to discriminate and identify challenging English vowel sounds. The efficacy of each perceptual task is assessed by comparing trainees to a group of untrained cohorts on their ability to identify and discriminate the target vowels in stimuli not present in training, namely in new nonword stimuli and real word stimuli produced by new talkers. In addition, the study also examines if the expected improvement in identification and discrimination as a result of HVPT is retained four months after the completion of the training regime. Based on previous research on HVPT, we expect that identification training will improve identification, and discrimination training will improve discrimination (e.g., Thomson, 2018). Further, given the categorical nature of the discrimination task used, we expect there will be cross-task effects and trainees will improve both the trained and the untrained ability (Flege, 1995; Wayland & Li, 2008; Shinohara & Iverson, 2018), although improvement in the trained ability may be greater due to procedural learning (Iverson et al., 2012; Nozawa, 2015). Finally, improvement is predicted to generalize to perception in new nonwords and in real words following previous studies that show an advantage of using nonword training stimuli, linked to a greater focus on phonetic form (Thomson & Derwing, 2016; Ortega et al., 2021; Carlet & Cebrian, 2022). Better overall performance with real words than with nonwords may be observed due to the role of lexical representations in L2 segmental perception (Yamada et al., 1997).

## Methodology

### *Participants*

Participants in this study were, initially, 44 Spanish/Catalan bilingual speakers (average age 19.4 years, 39 females), who were first-year students of English studies at a public university in Barcelona. Their exposure to English was mostly through their university classes as none had spent more than two months in an English-speaking country. No hearing problems were reported. The 44 participants were randomly distributed into two experimental groups and a control group (CG), although eventually, only 38 participants completed all the tests: 13 in the ID training group (IDG), 14 in the DIS training group (DISG), and 11 in CG. All groups were tested before training (pretest), after training (posttest), and four months after that (delayed posttest). Participants in CG were untrained from the pretest to the posttest, although they were given a combined DIS+ID training after the posttest and completed a second posttest afterward (posttest2). All participants received a small stipend.

### *Stimuli*

The focus of the study was the Southern Standard British English (SSBE) vowels /iː ɪ æ ʌ ɜː/, which are challenging for Catalan/Spanish learners of English, especially the /iː/-/ɪ/ and /æ/-/ʌ/ vowel contrasts (e.g., Cebrian, Gorba & Gavaldà, 2021; Mora et al., 2022). In the case of English /ɜː/, it was contrasted with two potentially confusable vowels, /ɛ/ and /ɑː/; therefore, within-trial contrasts in discrimination tasks and across-trial contrasts in identification tasks involved the /iː/-/ɪ/ and the /æ/-/ʌ/ vowel pairs, as well as the /ɛ/-/ɜː/ and the /ɑː/-/ɜː/ pairs. Thus, the stimuli consisted of monosyllabic CVC nonwords and real words containing the SSBE vowels /iː ɪ ɛ ɜː æ ʌ ɑː/, where the vowel was preceded and followed by an obstruent. The words were elicited from six talkers who were native speakers of SSBE and had spent most of their lives in the south of England (three females, three males, mean age: 27.8). None reported speaking any other languages fluently and/or

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

**Table 1.** Identification and discrimination training stimuli organized by vowel contrast

| /æ/ – /ʌ/ | | /iː/ – /ɪ/ | | /ɜː/ – /ɛ/ | | /ɜː/ – /ɑː/ | |
|---|---|---|---|---|---|---|---|
| zat | zut | jeet | jit | chert | chet | zert | zart |
| zad | zud | jeed | jid | cherd | ched | zerd | zard |
| vap | vup | veep | vip | jerp* | jep | jerp* | jarp |
| vab | vub | veeb | vib | jerb* | jeb | jerb* | jarb |
| vak | vuk | veek | vik | verk* | vek | verk* | vark |
| vag | vug | veeg | vig | verg* | veg | verg* | varg |
| dadge | dudge | deedge | didge | derge* | dedge | derge* | darge |
| tadge | tudge | teedge | tidge | terge* | tedge | terge* | targe |
| pav | puv | peedge | pidge | perf | peff | persh | parsh |
| bav | buv | beedge | bidge | berf | beff | bersh | barsh |
| kak | kuk | keedge | kidge | kerch* | ketch | kerch* | karch |
| gak | guk | geedge | gidge | gerch* | getch | gerch* | garch |

**Note*: Some /ɜː/ items appear twice as /ɜː/ was contrasted with /ɛ/ in half the trials and with /ɑː/ in the other half.

having any knowledge of Spanish and Catalan. Stimuli were embedded in a carrier sentence that facilitated the pronunciation of the nonwords (e.g., *It rhymes with badge: dagde. I say dadge now. I say dadge again*). All recordings took place in a soundproof chamber at a university in London, England, using Cool Edit 2000 software, a Rode Simply NT1-A microphone, and an Edirol UA-25 audio interface, and they were digitized at a 44.1 kHz sampling rate and 16-bit quantification. Two speakers (a male and a female) provided the testing stimuli (new nonwords and real words), and the remaining four (two male, two female) provided the training stimuli (all nonwords). Three native English speakers identified the selected stimuli accurately and consistently in an identification and goodness rating task.

The stimuli used for training were nonwords (e.g., *jeet, jit, dadge, dudge*; see Table 1 for the complete list of training stimuli). There were 12 words per vowel except for vowel /ɜː/, for which there were four additional words to have enough nonword CVC sequences containing this vowel that could be contrasted with /ɑː/ and /ɛ/ in discrimination training.

Testing stimuli, which were used in the pretest, posttest, posttest2, and delayed test, consisted of a new set of nonwords not used in training and a set of real words. Twenty-four real words and 24 nonwords were used in the discrimination test (four words per vowel except for /ɛ/ and /ɑː/, with two words each, contrasting with four /ɜː/ words). Twenty-six real words and nonwords were used in the identification tests, which were basically the same words used in the discrimination tests plus additional /ɑː/ and /ɛ/ words to obtain a balanced number of stimuli per vowel (see the procedure section and Appendix A).

### Procedure

The training was carried out by means of a seven-alternative forced-choice identification task (ID) and a categorical same/different AX discrimination task (DIS). Participants in both training regimes (IDG and DISG) were presented with the same number of stimuli; ID involved stimuli being presented individually whereas DIS presented stimuli in pairs. Thus, there were twice as many trials in each ID training session as in each DIS session. The control group was trained between posttest and posttest2 with three DIS sessions followed by three ID sessions. Testing involved the identification and discrimination of the target vowels in nonwords and real words. The pretest, posttest,

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

posttest2 (for CG) and delayed test were exactly the same. The participants also completed a perceptual assimilation task and a production task that are not reported in the current paper.

Training consisted of six 30-min sessions that took place over several weeks at a phonetics laboratory at a Spanish University. The software used was TP (Rauber, Rato, Kluge & Santos, 2011). In identification training (ID), a stimulus (nonword) was delivered through headphones at a comfortable sound level. Seven response options with a phonetic symbol and example words (i.e., /æ/ ash/mass, /ɑː/ arm/palm, /e/ less/west, /ɜː/ earth/first, /ɪ/ fish/his, /iː/ cheese/leaf, /ʌ/ sun/thus) were displayed on the screen. Due to restrictions of the TP software, some of the phonetic symbols were displayed with regular characters (i.e., /a:/ for /ɑː/, /3:/ for /ɜː/, /I/ for /ɪ/, /^/ for /ʌ/). Participants clicked on one of the seven options and received immediate feedback indicating the correct response. At the end of each session, participants were shown a global result (% correct answers). Each identification training session consisted of 480 trials, with a break after 240 trials. For each 240-trial section, there were 36 trials involving /iː ɪ æ ʌ ɑː ɜː/ and 24 trials involving /ɛ/ (a smaller number given that this vowel was not expected to pose a problem and was included to be contrasted with /ɜː/ in discrimination). Discrimination training (DIS) was implemented by means of a categorical AX discrimination (same/different) task in which participants had to indicate whether two given stimuli (produced by a female and a male speaker) contained the same or different vowels. Participants responded by clicking on one of two options (same or different) displayed on the screen. The order of the vowels and the talkers was counterbalanced throughout the tasks. There were 120 same-category and 120 different-category trials per DIS session. For each set of 120 different-category trials, there were 40 involving /æ/-/ʌ/, /ɪ/-/iː/ and 20 for /ɜː/-/ɛ/, /ɜː/-/ɑː/. Regarding the 120 same-category trials, there were 20 involving /æ/-/æ/, /ʌ/-/ʌ/, /iː/-/iː/, /ɪ/-/ɪ/, /ɜː/-/ɜː/ and 10 for /ɛ/-/ɛ, /ɑː/-/ɑː/. Immediate feedback was provided after each trial (correct or incorrect answer), and a global result was given at the end of the session.

Regarding the pre- and posttraining tests (posttest, posttest2 for CG, and delayed test), the ID tests included four words per vowel, each word produced by a male and a female speaker, and repeated twice (except for /ɛ/, which had fewer stimuli as explained above). The total number of trials was 104 (see Appendix A for a list of all the stimuli, number of talkers, repetitions, and total number of trials per test). The response alternatives used in ID testing were has/mass, palm/arch, send/mess, sir/earth, his/lift, cheese/leaf, and sun/thus. Some of these words were different from the options used in training, but they were equivalent in terms of syllabic structure and final consonants, which were different from the ones found in training words. The AX DIS task contained 96 trials (48 same-category and 48 different-category trials). Different-category trials consisted of four pairs of words for /iː/-/ɪ/ and /æ/-/ʌ/, and two pairs of words for /ɑː/-/ɜː/ and /ɛ/-/ɜː/. Each pair of words appeared four times to counterbalance the order of the vowels (V1-V2, V2-V1) and the talkers (T1-T2, T2-T1). Same-category trials consisted of four pairs of words for vowels /iː ɪ æ ʌ ɜː/ and two word pairs for vowels /ɑː/ and /ɛ/, and the order of talkers was also counterbalanced. The interstimulus interval was 1.15 s, long enough to prevent reliance on sensory memory and facilitate access to phonetic information stored in long-term memory (e.g., Højen and Flege, 2006). The DIS and ID tests were completed on the same day and were the only tasks completed that day. The order of the tests was the following: DIS real words, ID real words, DIS nonwords, ID nonwords. The first DIS and ID tests were preceded by a short practice session consisting of eight trials to familiarize participants with the task

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

and adjust the volume if necessary. Participants took between 25 and 35 min to complete all four tests. All tests (pretest, posttest, posttest2, and delayed test) were completed using Praat (Boersma & Weenink, 2018).

### *Data analysis*

The effects of the two training methods (ID and DIS training) on the identification and discrimination of English vowels presented in nonword and real word stimuli were examined by analyzing participants' results at pretest, posttest, posttest2 (for CG), and delayed test. Score (correctly or incorrectly identified or discriminated[1]) was the dependent variable. Two logistic mixed effects models were used (one for identification, and one for discrimination). Group (IDG, DISG, CG), test (pretest, posttest, posttest2, delayed test), word type (real word, nonword), and all possible two-way and three-way interactions were included as fixed factors. Subject-specific random intercepts and random slopes for time as well as word-specific random intercepts and slopes for talker were considered as random effects (Barr, Levy, Scheepers & Tily, 2013; Matuschek, Kliegl, Vasishth, Baayen & Bates, 2017). Random slopes were eliminated in the final model as the model did not converge in the case of the discrimination data, and so as to have comparable models for identification and discrimination. The difference between including or excluding random slopes in the case of the identification data was minimal and did not affect the levels of significance. Tukey's correction was used in pairwise comparisons. The analyses were performed using the GLIMMIX procedure of the SAS software (SAS Institute Inc., Cary, NC, USA). The significance level was set to 0.05. The results for identification are presented first, followed by the discrimination results.

## Results

### *Identification results*

The results for the pretest, posttest, posttest2 (for CG), and delayed test are presented in Table 2, which shows the mean % correct identification of the target English vowels in nonword and real word stimuli per group and test.[2] The mean identification accuracy scores and confidence intervals are graphically presented in Figure 1. The outcome of the logistic mixed effects model is given in Table 3. Test yielded a significant main effect ($p < .001$), explained by the general increase in identification accuracy from pretest (58%) to posttest (70.8%) and delayed test (76.5%), across groups and word type.[3] Identification scores were numerically higher in real words (74.4%) than in nonwords

[1]The discrimination results were also analyzed in terms of a sensitivity index (d-prime or d') based on signal detection theory (Macmillan & Creelman, 2005). A Spearman's correlation was then conducted to relate d' and discrimination accuracy values. The two measures were found to be strongly correlated ($\rho = .912$, $N = 174$, $p < .001$). For the sake of consistency with identification results and ease of interpretability, the analysis presented is based on discrimination accuracy scores, following some previous studies (e.g., Zhou, Dmitrieva & Olson, 2022).

[2]Due to space restrictions, and given the main goals of the study, results are analyzed globally and specific results for individual vowels or vowel pairs are not reported (see Carlet and Cebrian [2019] and Cebrian et al. [2021] for reported outcomes for similar populations).

[3]These mean values across groups include the results for CG, who had not undergone training yet at posttest. Excluding CG, the means for IDG and DISG combined are 56.6% at pretest, 73.4% at posttest, 74.7% at delayed test.

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

**Table 2.** Mean % correct identification and standard error per group and test for nonword and real word stimuli.

| | | Nonword stimuli | | Real word stimuli | |
|---|---|---|---|---|---|
| Group | Test | Mean | St. error | Mean | St. Error |
| ID | Pretest | 55.0 | 6.7 | 59.7 | 6.6 |
| | Posttest | 73.9 | 5.3 | 84.5 | 3.6 |
| | Delayed test | 76.0 | 5.1 | 83.9 | 3.8 |
| DIS | Pretest | 50.4 | 6.7 | 61.1 | 6.4 |
| | Posttest | 62.1 | 6.3 | 71.7 | 5.5 |
| | Delayed test | 63.3 | 6.3 | 74.5 | 5.2 |
| CTL | Pretest | 57.0 | 7.0 | 65.0 | 6.5 |
| | Posttest | 60.0 | 6.8 | 73.0 | 5.6 |
| | Posttest2 | 78.5 | 4.8 | 84.9 | 3.7 |
| | Delayed test | 75.6 | 5.6 | 85.5 | 3.8 |

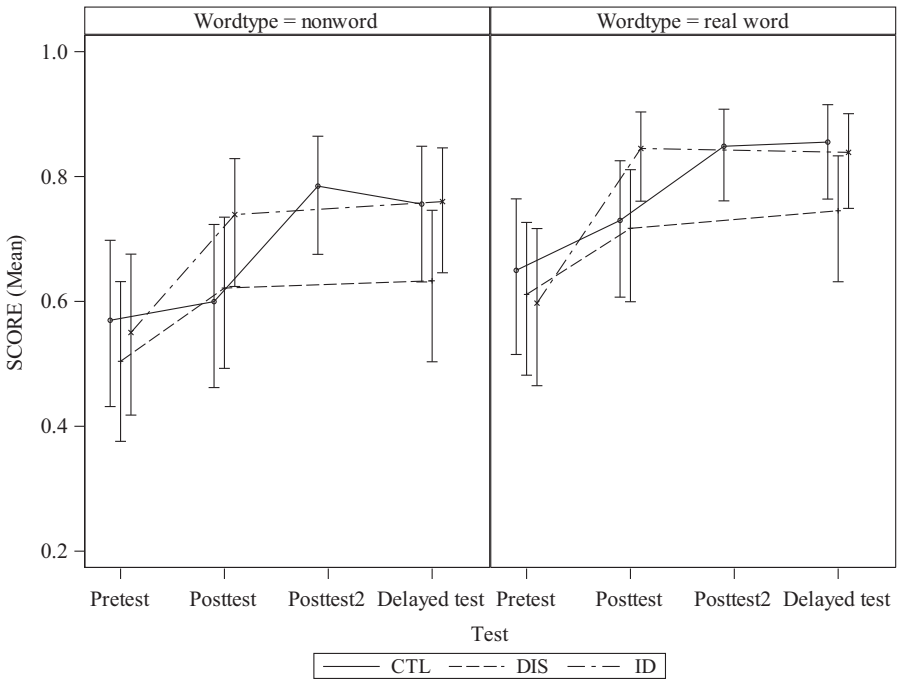


**Figure 1.** Line graphs with confidence intervals (whiskers) showing identification accuracy scores per group at pretest, posttest, posttest2 (CG), and delayed test for nonword and real words.

(65.2%), but the effect of word type did not reach significance. There was no effect of group but the interaction between test and group was significant ($p < .001$), as well as the interaction between test and word type ($p = .005$). No other interactions reached significance (see Table 3). Significant interactions were examined through pairwise comparisons with a Tukey correction (the results of all the pairwise comparisons are presented in Appendix B). Regarding the word type by test interaction, vowel identification in real word stimuli was significantly more accurate than in nonwords at posttest ($p = .047$) and the difference was marginally significant at delayed test ($p = .05$),

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

**Table 3.** Results of the logistic mixed model on identification accuracy

| Type III tests of fixed effects | | | | |
|---|---|---|---|---|
| Effect | Num. DF | Den. DF | F Value | Pr > F |
| Group | 2 | 23900 | 1.75 | 0.1734 |
| Test | 3 | 81 | 52.90 | <.0001 |
| Word type | 1 | 23900 | 2.80 | 0.0942 |
| Group*Test | 4 | 23900 | 7.04 | <.0001 |
| Group*Word type | 2 | 23900 | 0.41 | 0.6627 |
| Test*Word type | 3 | 23900 | 4.34 | 0.0046 |
| Group*Test*Word type | 4 | 23900 | 1.81 | 0.1240 |

but the two word types did not differ at pretest ($p = .251$) and posttest2 ($p = .149$). With respect to the test by group interaction, pairwise comparisons indicated that, for the trained groups, the difference between pretest and posttest results was significant at the $p < .001$ level. CG showed some nonsignificant improvement from the pretest to the posttest (5.7 percent points), which may be the result of continued exposure to the target language and familiarity with the task at the posttest. However, CG showed a much greater and significant improvement from posttest to posttest2 after undergoing training (15.1 percent points, from 66.8% to 81.9%), $p < .001$. The identification scores at the delayed test were significantly higher than at the pretest for all groups but did not differ from posttraining scores (posttest for IDG and DISG, posttest2 for CG), showing that the improvement from pre- to posttraining test was maintained at delayed test. The pairwise comparisons also showed that groups did not differ significantly at pretest (IDG: 57.4%, DISG: 55.8%, CG: 61.1%, across word type). At posttest, IDG's identification scores were significantly higher than those of the other two groups (IDG: 79.7%, DISG: 67.1%, CG: 66.8%), $p < .05$ in both cases. Finally, as an alternative way of exploring the test by group interaction, group results were compared in terms of the difference between the pretest and the posttest. IDG and DISG were compared on the amount of improvement from the pretest to the posttest, and the two trained groups together were compared to CG. The results indicated that IDG's improvement was greater than DISG's (22.3 and 11.3 percent points, respectively, $t = 3.29$, $p = .001$) and that trainees' improvement from pretest to posttest (IDG and DISG together) was significantly greater than CG's (16.8 and 5.7 percent points, respectively, $t = -3.15$, $p = .0016$). Therefore, ID training and DIS training resulted in a significant improvement in identification accuracy, in contrast to the lack of significant improvement for CG, and the improvement was greater for IDG than for DISG, and for the trained groups (IDG and DISG together) than for CG. These outcomes are revisited in the discussion section in light of the study's predictions. The results of the discrimination tests are presented next.

### *Discrimination results*

Table 4 presents the % correct discrimination per group at pretest, posttest, posttest2 for CG, and delayed test, for nonword and real word stimuli (see footnote 1). The results are presented graphically in Figure 2, which includes confidence intervals. Table 5 shows the outcome of the logistic mixed effects model, which mirrors the results obtained in the identification test, with the addition of a significant effect of word type. Thus, test yielded a significant main effect ($p < .001$), reflecting the increase in correct

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

**Table 4.** Mean % correct discrimination and standard error per group and test for nonword and real word stimuli

| | | Nonword stimuli | | Real word stimuli | |
|---|---|---|---|---|---|
| Group | Test | Mean | St. error | Mean | St. error |
| ID | Pretest | 72.9 | 3.7 | 75.4 | 3.5 |
| | Posttest | 80.0 | 3.0 | 85.4 | 2.4 |
| | Delayed test | 79.1 | 3.2 | 85.1 | 2.5 |
| DIS | Pretest | 73.3 | 3.6 | 75.3 | 3.4 |
| | Posttest | 79.5 | 3.0 | 84.6 | 2.4 |
| | Delayed test | 79.6 | 3.1 | 87.7 | 2.1 |
| CTL | Pretest | 71.9 | 4.0 | 80.5 | 3.1 |
| | Posttest | 74.5 | 3.7 | 82.6 | 2.9 |
| | Posttest2 | 79.7 | 3.2 | 86.4 | 2.4 |
| | Delayed test | 81.4 | 3.3 | 88.5 | 2.3 |

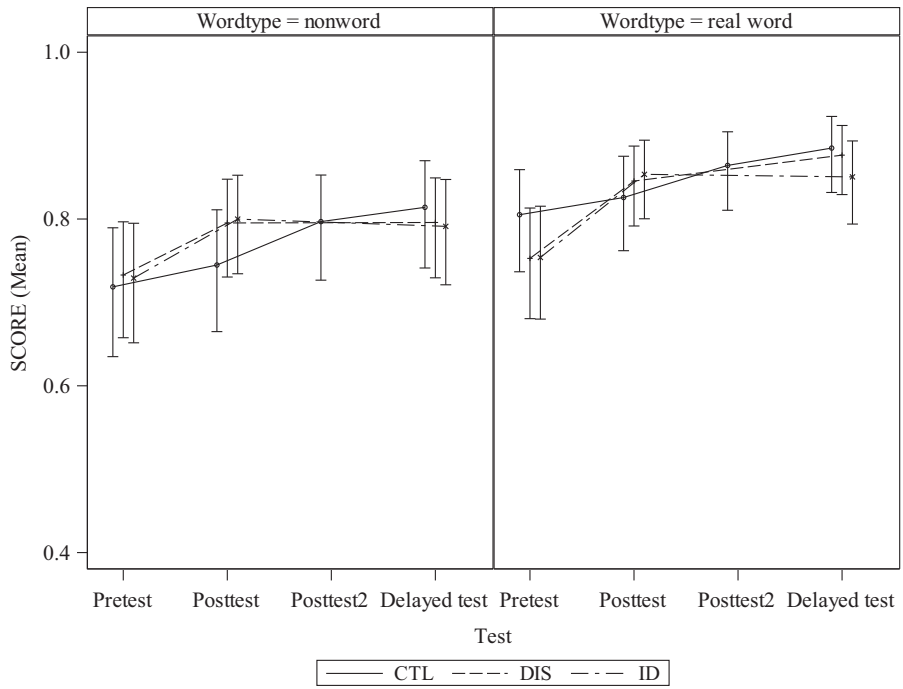


**Figure 2.** Line graphs with confidence intervals (whiskers) showing discrimination accuracy scores per group at pretest, posttest, posttest2 (CG), and delayed test for nonword and real words.

discrimination accuracy from pretest (74.9%) to posttest (81.1%) and delayed test (83.5%), across groups and word type.[4] Discrimination was more accurate in real words (83.1%) than in nonwords (77.2%), reaching significance at the $p < .05$ level. Group did

[4]Excluding CG, the mean scores for IDG and DISG combined were 74.2% at pretest, 82.8% at posttest, 82.3% at delayed test.

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

**Table 5.** Results of the logistic mixed model on discrimination accuracy

| Type III tests of fixed effects | | | | |
|---|---|---|---|---|
| Effect | Num DF | Den DF | F Value | Pr > F |
| Group | 2 | 22052 | 0.02 | 0.9841 |
| Test | 3 | 81 | 27.85 | <.0001 |
| Word type | 1 | 22052 | 5.56 | 0.0184 |
| Group*Test | 4 | 22052 | 2.72 | 0.0280 |
| Group*Word type | 2 | 22052 | 2.52 | 0.0806 |
| Test*Word type | 3 | 22052 | 3.34 | 0.0184 |
| Group*Test*Word type | 4 | 22052 | 1.03 | 0.3892 |

not yield a significant main effect, but the test by group interaction was significant ($p$ = .028), and there was also a significant interaction between test and word type ($p$ = .018; see Table 5 for details). The results of all the Tukey-corrected pairwise comparisons exploring these interactions are given in Appendix B. With respect to the test by word type interaction, real words obtained significantly higher accuracy scores than nonwords at posttest ($p$ = .019), posttest2 ($p$ = .016), and delayed test ($p$ = .003), but not at pretest ($p$ = .161). The interaction between test and group is explained by several facts. First, as was found for identification, the trained groups' scores at the posttest were significantly higher than at the pretest (IDG: pretest = 74.2%, posttest = 82.8%; DISG: pretest = 74.3%, posttest = 82.2%; $p$ < .001 in both cases), but there was no significant difference between pre- and posttest results for CG (pretest = 76.5%, posttest = 78.8%, $p$ = .604). CG's scores improved significantly after training (posttest2 = 83.3%), $p$ = .043. On the other hand, the results of the delayed test did not differ from posttraining scores, showing that the improvement was maintained four months after training had ended for all three groups. Between-group comparisons showed that there was no significant difference between any groups at any test time (see Table B9 in Appendix B). As was done for the identification results, the test by group interaction was explored further by comparing group results in terms of the difference between pretest and posttest. Again, the two trained groups together (IDG+DISG) were compared to CG, and IDG and DISG were also compared. The results revealed that the trainees' improvement from pretest to posttest (IDG and DISG together) was significantly greater than CG's (8.2 and 2.4 percent points, respectively, $t$ = -2.75, $p$ = .0061) and that DISG's and IDG's improvement did not differ significantly (8.6 and 7.8 percent points, respectively, $t$ = 0.37, $p$ = .711). In brief, the results show that IDG and DISG, but not CG, improved significantly from pretest to posttest and that the trained groups outperformed CG, but did not differ from one another.

Finally, to examine if the degree of improvement in discrimination and identification were related at an individual level, a Pearson's correlation was conducted involving each individual's improvement in each type of task. Specifically, the difference between pre- and posttraining tests in percent points was calculated for each participant in each task across nonword and real-word stimuli. For IDG and DISG, improvement reflects the difference between posttest and pretest, while for CG the difference between posttest2 and posttest was calculated. The results indicated that improvement in the two measures was significantly correlated, $r$ = .324, $N$ = 76, $p$ = .004, as illustrated by the scatterplot in Figure 3.

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

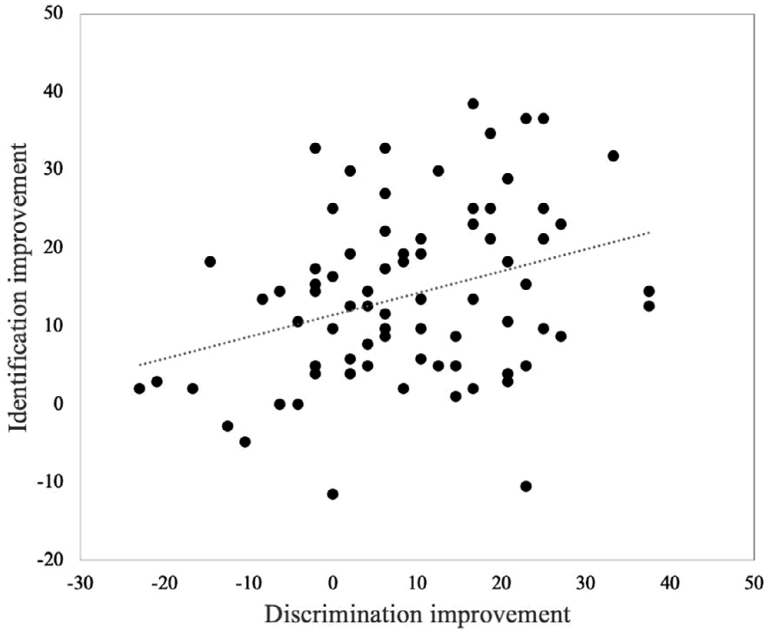


**Figure 3.** Scatterplot of identification (*y*-axis) and discrimination (*x*-axis) improvement from pre- to posttraining test (in percent points) per individual across real and nonword stimuli.

## Discussion

### *Identification and discrimination of L2 vowels*

The first goal of this paper was to compare the effect of ID and DIS training on both the discrimination and the identification of L2 vowels. In accordance with our predictions, both ID and DIS training had a positive effect, as shown by the significant differences between the pretest and posttest for IDG and DISG in both identification and discrimination accuracy, and the fact that the trained groups outperformed CG in both perceptual tasks in terms of rate of improvement. The control group improved numerically from the pretest to the posttest but not significantly. Recall that pre- and posttests included new nonwords and real words and thus the improvement from pre- to posttest constitutes a measure of generalization, as discussed below. Thus, the first finding of the current study is the fact that both training methods (ID and categorical DIS) were effective in enhancing not only the object of training (ID for IDG and DIS for DISG) but also the untrained tasks (DIS for IDG and ID for DISG). These results are in line with previous research reporting that ID can improve categorical discrimination (Iverson et al., 2012; Wayland & Li, 2008), and categorical DIS improves identification (Flege, 1995; Nozawa, 2015; Carlet & Cebrian, 2022), and with a previous study that compared identification training with a discrimination training that included both auditory and categorical tasks and found reciprocal effects (Shinohara & Iverson, 2018). Studies that report no effect of DIS on either identification or discrimination (e.g., Strange & Dittman's (1984) lack of generalization effects), or of ID on discrimination (Lengeris & Hazan, 2010), made use of DIS tasks that involved synthetic stimuli and low interstimulus intervals. Thus, no cross-task improvement was found when training relied on sensory information and low-level processing. In this sense, the results of

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

earlier studies support the idea that when listeners perform the same/different discrimination task, they rely on short-lived sensory information that is useful for determining if two stimuli are physically identical or not but is less conducive to the development of long-term memory representations for L2 categories (Flege, 1995). Identification training, particularly when using multiple stimuli from the same category, increases listeners' sensitivities to the common properties that make a given category distinguishable from other categories and thus promotes the formation of more robust long-term memory representations of those categories (Jamieson & Morosan, 1986). In contrast to an auditory DIS task that relies on sensory information (e.g., Strange & Dittman, 1984), a categorical DIS task includes multiple tokens from the same category (e.g., from different talkers or different productions by the same talker), and involves evaluating if the stimuli presented are sufficiently different to belong to separate categories. Thus, it is possible that when performing a categorical DIS task, listeners may in fact be identifying each stimulus in the pair individually before determining if they are the same or different sounds. This possibility was suggested by Shinohara and Iverson (2018), the only previous study to our knowledge to have contrasted the effect of ID and DIS training (using a combination of auditory and discrimination tasks) on both measures in a single study and whose results coincide with the current results. These authors explained that DIS trainees may covertly label the phonemes when performing a categorical discrimination task. Recall that in a categorical DIS task same trials include two physically different stimuli from the same category. Thus, precisely given that sameness cannot be judged on the basis of physical identity and that stimuli may be identified prior to being compared, a categorical DIS task may encourage a similar level of phonological encoding to that of ID tasks. As previously discussed, ID involves determining to which of a set of internal representations a given stimulus belongs and entails a higher-level phonological encoding (Flege, 1995; Logan & Pruitt, 1995). Hence both ID and categorical DIS tasks may involve similar levels of processing that enhance the formation of more robust L2 categories (Polka, 1992; Flege, 1995; Wayland & Li, 2008). An alternative view may also be considered, that is, the possibility that training identification may enhance discrimination since identifying a given vowel category correctly implies distinguishing it from perceptually close vowels (i.e., tokens of /æ/ and /ʌ/ stimuli cannot be identified correctly unless the listener perceives these two vowels as different). Therefore, ID trainees may have been implicitly trained on discrimination by being presented with confusable categories such as /æ/ and /ʌ/ across trials in the ID task, which can explain why ID training is as effective as DIS training in improving discrimination.[5] In any event, the significant correlation between identification and discrimination improvement found in the current study (also reported by Shinohara and Iverson[2018]) supports the idea that ID and categorical DIS tasks may involve comparable strategies on the part of the listener and may enhance similar abilities and be mutually beneficial.

The second main finding of the study is that while both IDG and DISG experienced a very similar improvement in discrimination as a result of their respective training regimes, IDG clearly outperformed DISG in identification accuracy after training. This seems to point to an asymmetry between the two tasks, as ID training was found to improve identification accuracy more than DIS training, but DIS did not improve discrimination more than ID. This outcome was not expected, as we predicted similar cross-task effects given the assumption that both tasks involve similar processes.

[5]We thank an anonymous reviewer for pointing out this possibility.

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

The better results for IDG with identification could be partly explained by procedural learning, that is, the result of task familiarity, as IDG outperformed DISG precisely in identification. However, if differences between the two methods were simply explained by task familiarity, we would also expect DISG to outperform IDG in the discrimination results, and this was not the case. The current results are in fact in agreement with previous research that has compared ID and DIS training within the same study that has shown an advantage of ID for vowel identification (Nozawa, 2015; Carlet & Cebrian, 2022). ID's superiority in identification may stem from methodological differences between the two tasks. Recall that while the two training regimes used the exact same number and set of stimulus words, ID involved twice the number of trials, as a single word was presented at a time. Thus, given that feedback was provided after each trial, ID offered twice the amount of feedback, and of a more specific kind (the vowel identity). Even if categorical DIS may involve some level of identification to assess the shared identity or not of the two physically different stimuli, the DIS task itself does not consist of labeling the stimuli using one of several options presented. Hence, the greater number of trials and consequently greater opportunities for feedback, together with the use of a more explicit type of feedback in ID training, may account for ID's greater benefit on identification over DIS training.

Regarding the control group, the nonsignificant numerical improvement observed from the pretest to the posttest may be attributed to task familiarity as well as to continuous exposure to the target language as participants were undergraduate students majoring in English. CG's main and significant improvement occurred after undergoing DIS+ID training, reaching accuracy levels comparable to those of IDG in identification and to both groups in discrimination. This would indicate that combined use of tasks may be as efficient as the use of ID alone, although to fully assess this possibility a different study design would be necessary with DIS+ID training and ID training being implemented in parallel.

### *Measures of robust learning: generalization and retention*

The results of the posttests indicate that both types of perceptual tasks promoted generalization of learning as testing stimuli involved new voices and new words (recall that tests included stimuli and talkers not heard during training). Thus, training with nonwords resulted in an improvement of vowel perception in untrained nonwords as well as real words. This finding is in line with our predictions based on recent studies suggesting that the use of nonwords allows focusing on the phonetic form and avoids word familiarity effects as well as lexical and orthographic biases (Thomson & Derwing, 2016; Fouz-González & Mompeán, 2021; Ortega et al., 2021). This potential advantage of nonword stimuli, however, may not be unconstrained. Mora et al. (2022) reported that the advantage of training with nonwords over real words disappeared when background noise was added: the use of masking noise in an immediate repetition task in production training revealed a detrimental effect of noise with nonword stimuli but not with real word stimuli. According to the authors, the presence of noise hinders the focus on phonetic form, which is precisely what makes nonword stimuli advantageous.

On the other hand, participants were more successful in identifying and discriminating target vowels in real words than in nonwords (74.4% identification accuracy with real words vs. 65.2% for nonwords, and 83.1% discrimination accuracy with real words vs. 77.2% with nonwords, across groups and tests) although the difference reached statistical significance only with discrimination. This general real-word advantage in L2 vowel perception was also expected in light of earlier findings that show better

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

perception of L2 sounds in real words than in nonwords (Mora, 2005; Rato & Carlet, 2020) and that suggest that word knowledge and lexical representations play a role in L2 segmental perception (e.g., Yamada et al., 1997). Nevertheless, given that all stimuli were presented in isolation, and that real word stimuli involved minimal pairs (e.g., *bead*, *bid*, *bed*, *bad*, *bud*, *bard*, *bird*, see Appendix A), it remains to be assessed how lexical status may be an advantage with words that are likely phonetically confusable and possibly stored with ambiguous or neutralized lexical representations (Darcy & Holliday, 2019). Exploring the relationship between phonological and lexical representations lies beyond the scope of the present paper, but the finding that learning acquired through training with phonetically-oriented stimuli (nonwords) transfers to real word perception underscores the efficacy of the training methodology.

Regarding long-term retention, the results of a delayed test, completed four months after training ended, consistently replicated the posttraining results for all groups, with mean correct identification and discrimination accuracy values that were always very close and never differed significantly from posttraining scores. These results show evidence of long-term retention after a longer period than most previous studies (generally up to two or three months, see Thomson, 2018). Lively et al. (1994) reported that Japanese learners of English retained the improvement in their identification of the /r/-/l/ contrast three months after training, but signs of decline were observed after six months. Iverson and Evans (2009) found that L2 learners were able to retain their improvement in vowel identification an average of four months after training. Longer retention has been reported in a study testing American English speakers' perception of Mandarin tones, where improvement was still evident six months after the posttest (Wang, Spence, Jongman & Sereno, 1999). Longer retention may be more likely with suprasegmental phenomena like tone than with segmental contrasts, an issue that remains to be explored (Thomson, 2018). The evidence of both generalization and retention of learning indicates that both ID and DIS training can successfully trigger the development of L2 categories that are robust enough to perceive L2 sounds accurately disregarding interstimulus variations resulting from talker or speech rate differences that are irrelevant to category identity (Flege, 1995). Importantly, retention and generalization were not only found for the perceptual task that was the object of training (ID for IDG and DIS for DISG) but equally for both perceptual tasks for all groups. The current results did not show a tendency for DIS trainees to show more evidence of retention than ID trainees in identification, unlike a couple of previous studies (Flege, 1995; Carlet & Cebrian, 2022). Flege (1995) found that discrimination trainees' scores in ID accuracy at delayed tests had in fact increased while ID trainees' scores had decreased a little. In addition, DIS training, but not ID training, was found to improve the identification of sounds that are present in the stimuli but not the focus of training (Carlet & Cebrian, 2022). These outcomes were interpreted to indicate that DIS training is more efficient at consolidating learning. The current study did not find such an advantage for DIS training, but it differs from previous studies in several ways. It tested both ID and DIS, not only ID, and focused on vowels (as opposed to consonants or both vowels and consonants); in addition, in the present study the delayed test was administered at a later time than in those previous studies (four months vs. two months after the posttest).

## Final conclusions, limitations, and further research

The current study contrasted the effect of two perceptual training approaches (ID and DIS) on both L2 vowel identification and L2 vowel discrimination within the same

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

study and included measures of robust learning (generalization and retention), thus allowing a thorough examination of the efficacy of the two methods. The results provide evidence of the suitability of both ID and DIS training for improving both the identification and the discrimination of L2 vowels. This finding is supported by strong generalization and retention effects, as training with nonwords generalized to new nonwords and real words, and to new voices, and retention of learning was evident four months after training had finished. ID appeared to be more successful than DIS in improving vowel identification accuracy, a finding that is in agreement with some previous research (Nozawa, 2015; Carlet & Cebrian, 2022). The cross-task effects are explained by the fact that identification and categorical discrimination may involve similar levels of processing given the presence of multiple stimuli, the long interstimulus interval used in DIS and, consequently, the likelihood that categorical DIS involves the identification of each member in the stimulus. On the other hand, the fact that both ID and DIS equally improve discrimination but ID training has a greater effect on identification than DIS training can be explained by crucial methodological differences between the two training regimes, namely the type and the amount of feedback obtained with identification. The current study contributes to the line of recent studies showing the efficacy of using nonword stimuli in HVPT for the learning of segmental contrasts, further supported by the transfer of this knowledge to real word stimuli.

The present study had a few limitations. First, although a total of 44 participants were recruited, the fact that participants were distributed among three groups and that only 38 completed all the tasks, resulted in a relatively small sample for each training method (13 in IDG, 14 in DISG, and 11 in CG). Even if significant effects of training were obtained, group differences were not always observed, which could have emerged with a larger sample. In addition, the study was limited to a subset of L2 vowels, which included challenging contrasts for the population under study, but it did not examine the whole vowel inventory, nor did it explore the identification and discrimination of consonant sounds. Regarding the training stimuli, the voicing of the final consonant was not completely controlled, as there were 40 words ending in a voiced obstruent and 48 in a voiceless obstruent. English vowels are known to be shorter preceding a voiceless consonant, which may have made vowels before a voiced obstruent easier to perceive. We expect that the impact of this design problem on the overall training regime to have been minimal given the small difference in the number of tokens per voicing condition, and testing stimuli were appropriately balanced, but future research should address this and the previously mentioned limitations.

The current study adds to the wealth of research that generally supports the effectiveness of HVPT. These studies generally provide evidence of what can be achieved through phonetic training, but the question that remains is how training specifically affects the process of L2 category formation (Iverson & Evans, 2009; Shinohara & Iverson, 2018). In other words, it is unclear if improvement from pre- to posttraining actually reflects real changes in L2 categorization. It has been proposed that training may help learners to be more consistent and successful in using their existing categories to perceive their L2 sounds without necessarily altering the learners' internal representations of L2 categories (Iverson, Hazan & Bannister, 2005; Shinohara & Iverson, 2018). Iverson and colleagues found that Japanese learners of English became more consistent at using not only the primary acoustic cue (F3) but also an irrelevant or secondary cue for native speakers (F2) in their perception of the English /r/-/l/ contrast (e.g., Shinohara & Iverson, 2021). This is in line with Polka's (1992) observation that learners who undergo identification training

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

may learn to identify L2 sounds accurately by paying attention to characteristics that may help differentiate nonnative categories, but which may not be the properties attended to by native speakers. More research is needed to fully evaluate what truly changes as a result of phonetic training and to investigate if and how internal representations of L2 categories can be altered through phonetic training. For instance, some recent research points to the use of tasks aimed at changing cue-weighting, e.g., through cue enhancement or exaggeration, as possible inducers of actual category changes (Zhang et al., 2021b).

Finally, another aspect to be considered is the pedagogical potential of HVPT for pronunciation teaching and learning (Thomson, 2011). The current study shows that ID and categorical DIS tasks are successful methods of improving L2 vowel perception, although ID may be more suitable for training identification. Logan and Pruitt (1995) indicate that categorical DIS tasks might be more effective than identification tasks in the early stages of learning, when identification labels may not be fully understood and reliable. Carlet (2017) and Shinohara and Iverson (2018) suggest that a combination of both ID and DIS tasks might be beneficial as they may add variation and flexibility to the training regimes. In fact, a complete evaluation of training methods should also consider the learners' reactions to the training methods. Flege (1995) reported that ID trainees felt that training was more enjoyable, interesting, and beneficial than DIS trainees did. Similar impressions are reported by Carlet (2017). Possibly, a full examination of training methodologies should take into consideration not only the objective efficacy of the method but also the subjective impressions of the learners undergoing training.

**Supplementary material.** The supplementary material for this article can be found at https://doi.org/10.1017/S0272263124000408.

**Acknowledgements.** This work was supported by Research Grant Nos. FFI2017-88016-P and PID2021-122396NB- I00 to the first author, No. PID2019-107814GB-I00 from the Spanish Ministries of Economy and Competitiveness and Science and Innovation, and research grant 2021SGR00544 to the Experimental Phonetics research group from the Catalan Agency for Management of University and Research Grants (AGAUR).

**Competing interest.** The authors declare none.

## References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*, 255–278. doi: 10.1016/j.jml.2012.11.001

Boersma, P., & Weenink, D. (2018). "Praat: Doing phonetics by computer" [Computer program], version 6.0.43, retrieved from http://www.praat.org/

Brekelmans, G., Lavan, N., Saito, H., Clayards, M., & Wonnacott, E. (2022). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *Journal of Memory and Language*, *126*, 104352. https://doi.org/10.1016/j.jml.2022.104352

Carlet, A. (2017). *L2 perception and production of English consonants and vowels by Catalan speakers: The effects of attention and training task in a cross-training study*. [Unpublished doctoral dissertation]. Universitat Autònoma de Barcelona.

Carlet, A., & Cebrian, J. (2019). Assessing the effect of perceptual training on L2 vowel identification, generalization and long-term effects. In A. M. Nyvad, M. Hejná, A. Højen, A. B. Jespersen, & M. H. Sørensen (Eds.), *A Sound Approach to Language Matters–In Honor of Ocke-Schwen Bohn* (pp. 91–119). Dept. of English, School of Communication & Culture, Aarhus University.

Carlet, A., & Cebrian, J. (2022). The roles of task, segment type, and attention in L2 perceptual training. *Applied Psycholinguistics*, *43*(2), 271–299. https://doi.org/10.1017/s0142716421000515

Cebrian, J., Gorba, C., & Gavaldà, N. (2021). When the easy becomes difficult: Factors affecting the acquisition of the English /iː/-/ɪ/ contrast. *Frontiers in Communication*, *6*, 1–17. https://doi.org/10.3389/fcomm.2021.660917

Darcy, I., & Holliday, J. J., (2019). Teaching an Old Word New Tricks: Phonological Updates in the L2 Mental Lexicon. *Pronunciation in Second Language Learning and Teaching Proceedings 10*(1).

Flege, J.E. (1995). Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics*, *16*, 425–442. https://doi.org/10.1017/S0142716400066029

Fouz-González, J., & Mompean, J. A. (2021). Phonetic symbols vs keywords in perceptual training: the learners' views. *ELT Journal*, *75*(4), 460–470. https://doi.org/10.1093/elt/ccab037

Georgiou, G. P. (2021). Effects of phonetic training on the discrimination of second language sounds by learners with naturalistic access to the second language. *Journal of Psycholinguistic Research*, *50*(3), 707–721.https://doi.org/10.1007/s10936-021-09774-3

Højen, A., & Flege, J. E. (2006). Early learners' discrimination of second-language vowels. *Journal of the Acoustical Society of America*, *119*(5), 3072–3084. doi:10.1121/1.2184289

Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *The Journal of the Acoustical Society of America*, *126*(2), 866–877. https://doi.org/10.1121/1.3148196

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults, *Journal of the Acoustical Society of America*, *118*, 3267–3278.

Iverson, P., Pinet, M., & Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(01), 145–160. https://doi.org/10.1017/s0142716411000300

Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones. *Perception & Psychophysics*, *40*(4), 205–215. https://doi.org/10.3758/bf03211500

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, *26*(2), 227–247.

Law, I. L. G., Grenon, I., Sheppard, C., & Archibald, J. (2019). Which is better: Identification or discrimination training for the acquisition of an English coda contrast. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 939–943). Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Lengeris, A., & Hazan, V. (2010). The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of English vowels for native speakers of Greek. *The Journal of the Acoustical Society of America*, *128*(6), 3757–3768. https://doi.org/10.1121/1.3506351

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3), 1242–1255.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y. I., & Yamada, T. (1994). Training Japanese listeners to identify English /r/and/l/ III: Long-term retention of new phonetic categories. *The Journal of the Acoustical Society of America*, *96*(4), 2076–2087.

Logan, J., & Pruitt, J. (1995). Methodological issues in training listeners to perceive non-native phonemes. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross Language Research* (pp. 351–378). York Press.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Lawrence Erlbaum Associates Publishers.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of memory and language*, *94*, 305–315.

Mora, J. C. (2005) Lexical knowledge effects on the discrimination of non-native phonemic contrasts in words and nonwords by Catalan/Spanish bilingual learners of English. In Hazan, V. and Iverson, P. (eds) *Proceedings of the ISCA Workshop on Plasticity in Speech Perception* (pp. 43–46). London: Dept. of Phonetics and Linguistics, University College London.

Mora, J.C., Ortega, M., Mora-Plaza, I., & Aliaga-García, C. (2022) Training the pronunciation of L2 vowels under different conditions: the use of non-lexical materials and masking noise. *Phonetica*, *79*(1), 1–43. https://doi.org/10.1515/phon-2022–2018

Nozawa, T. (2015). Effects of training methods and attention on the identification and discrimination of American English coda nasals by native Japanese listeners. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. The University of Glasgow. ISBN 978-0-85261-941-4.

Ortega, M., Mora-Plaza, I., & Mora, J. (2021) Differential effects of lexical and non-lexical high-variability phonetic training on the production of L2 vowels. In Kirkova-Naskova, A., Henderson, A., & Fouz-González, J. (Eds.) *English Pronunciation Instruction: Research-based Insights* (pp. 328–355). John Benjamins.

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*(1), 461–472.

Polka, L. (1992). Characterizing the influence of native language experience on adult speech perception. *Perception & Psychophysics*, *52*(1), 37–52. https://doi.org/10.3758/bf03206758

Rato, A., & Carlet, A. (2020). Second language perception of English vowels by Portuguese learners: The effect of stimulus type. *Ilha do Desterro*, *73*, 205–226.

Rauber, A. S., Rato, A., Kluge, D., & Santos, G. R. (2011). TP software.

Shinohara, Y., & Iverson, P. (2018). High variability identification and discrimination training for Japanese speakers learning English /r/–/l/. *Journal of Phonetics*, *66*, 242–251. https://doi.org/10.1016/j.wocn.2017.11.002

Shinohara, Y., & Iverson, P. (2021). The effect of age on English/r/-/l/perceptual training outcomes for Japanese speakers. *Journal of Phonetics*, *89*, 101108

Strange, W. (1992). Learning non-native phoneme contrasts: Interactions among subject, stimulus, and task variables. In Y. Tohkura, E. Vatikitois-Bateson, and Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure* (197–220). Tokyo: Ohm.

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psycophysics*, *36*(2), 131–145. https://doi.org/10.3758/bf03202673

Thomson, R. I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *CALICO Journal*, *28*(3), 744–765. https://doi.org/10.11139/cj.28.3.744–765

Thomson, R. I. (2018). High variability [pronunciation] training (HVPT): A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation*, *4*(2), 208–231. https://doi.org/10.1075/jslp.17038.tho

Thomson, R. I., & Derwing, T. M. (2016). Is phonemic training using nonsense or real words more effective? In J. Levis, H. Le., I. Lucic, E. Simpson and S. Vo (Eds.), *Proceedings of the 7thPronunciation in Second Language Learning and Teaching Conference* (pp. 88–97). Ames, IA: Iowa State University.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, *106*(6), 3649–3658.

Wayland, R. P., & Li, B. (2008). Effects of two training procedures in cross-language perception of tones. *Journal of Phonetics*, *36*(2), 250–267. https://doi.org/10.1016/j.wocn.2007.06.004

Wee, D. T., Grenon, I., Sheppard, C., & Archibald, J. (2019). Identification and discrimination training yield comparable results for contrasting vowels. In Calhoun, S., Escudero, P., Tabain, M., & Warren, P., (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 939–943). Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Yamada, R. A., Tohkura, Y., & Kobayashi, N. (1997). Effect of word familiarity on non-native phoneme perception: Identification of English /r/, /l/, and /w/ by native speakers of Japanese. In James, A. and Leather, J. (eds.) *Second language speech. Structure and process* (pp. 103–117). Berlin: Mouton de Gruyter.

Zhang, X., Cheng, B., & Zhang, Y. (2021a). The role of talker variability in nonnative phonetic learning: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, *64*(12), 4802–4825.

Zhang, X., Cheng, B., Qin, D., & Zhang, Y. (2021b). Is talker variability a critical component of effective phonetic training for nonnative speech? *Journal of Phonetics 87*, 101071. https://doi.org/10.1016/j.wocn.2021.101071

Zhou, A., Dmitrieva, O., & Olson, D. J. (2022). The effect of allophonic variability on L2 contrast perception: Evidence from perception of English vowels. *JASA Express Letters*, *2*(12), 125201. https://doi.org/10.1121/10.0016602

## Appendix A Testing stimuli

**Table A1.** Identification test. Real word stimuli

| Vowel | Real words | | | | Talkers | Repetitions | Trials |
|---|---|---|---|---|---|---|---|
| /iː/ | bead | lead | beat | feet | 2 | 2 | 16 |
| /ɪ/ | bid | lid | bit | fit | 2 | 2 | 16 |
| /æ/ | bad | bag | back | cat | 2 | 2 | 16 |
| /ʌ/ | bud | bug | buck | cut | 2 | 2 | 16 |
| /ɜː/ | bird | berg | berk | hurt | 2 | 2 | 16 |
| /ɑː/ | bard ×2 | bark | heart | | 2 | 2 | 16 |
| /ɛ/ | bed | beg | pet ×2 | | 2 | 1 | 8 |
| | | | | | | Total number of trials: 104 | |

*Note:* Two different productions per talker for *bard* and *pet* were used.

**Table A2.** Identification test. Nonword stimuli

| Vowel | Nonwords | | | | Talkers | Repetitions | Trials |
|---|---|---|---|---|---|---|---|
| /iː/ | theeb | teeve | feep | beesh | 2 | 2 | 16 |
| /ɪ/ | thib | tiv | fip | bish | 2 | 2 | 16 |
| /æ/ | jad | gadge | jat | dap | 2 | 2 | 16 |
| /ʌ/ | jud | gudge | jut | dup | 2 | 2 | 16 |
| /ɜː/ | ferb | derve | jert | terch | 2 | 2 | 16 |
| /ɑː/ | farb | darve | jart | barsh | 2 | 2 | 16 |
| /ɛ/ | jed | tech | | | 2 | 2 | 8 |
| | | | | | | Total number of trials: 104 | |

**Table A3.** Discrimination test. Real word stimuli

| | Target pair | Real words | | | | Pairs | Order* | Trials |
|---|---|---|---|---|---|---|---|---|
| Different pairs | /iː/-/ɪ/ | bead-bid | lead-lid | beat-bit | feet-fit | 4 | 4 | 16 |
| | /æ/-/ʌ/ | bad-bud | bag-bug | back-buck | cat-cut | 4 | 4 | 16 |
| | /ɜː/-/ɑː/ | bird-bard | hurt-heart | | | 2 | 4 | 8 |
| | /ɜː/-/ɛ/ | berg-beg | pert-pet | | | 2 | 4 | 8 |
| | /iː/-/iː/ | bead-bead | lead-lead | beat-beat | feet-feet | 4 | 2 | 8 |
| Same pairs | /ɪ/-/ɪ/ | bid-bid | lid-lid | bit-bit | fit-fit | 4 | 2 | 8 |
| | /æ/-/æ/ | bad-bad | bag-bag | back-back | cat-cat | 4 | 2 | 8 |
| | /ʌ/-/ʌ/ | bud-bud | bug-bug | buck-buck | cut-cut | 4 | 2 | 8 |
| | /ɜː/-/ɜː/ | bird-bird | berg-berg | pert-pert | hurt-hurt | 4 | 2 | 8 |
| | /ɑː/-/ɑː/ | bard-bard | heart-heart | | | 2 | 2 | 4 |
| | /ɛ/-/ɛ/ | beg-beg | pet-pet | | | 2 | 2 | 4 |
| | | | | | | Total number of trials: 96 (48 same and 48 different) | | |

*Note*:
*Four possible combinations: two talker orders T1-T2, T2-T1, and two vowel orders: V1-V2, V2-V1.

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

**Table A4.** Discrimination test. Nonword stimuli

| | Target pair | Nonwords | | | | Pairs | Order* | Trials |
|---|---|---|---|---|---|---|---|---|
| Different pairs | /iː/-/ɪ/ | theeb-thib | teeve-tiv | feep-fip | beesh-bish | 4 | 4 | 16 |
| | /æ/-/ʌ/ | jad-jud | gadge-dudge | jat-jut | dap-dup | 4 | 4 | 16 |
| | /ɜː/-/ɑː/ | derve-darve | jert-jart | | | 2 | 4 | 8 |
| | /ɜː/-/ɛ/ | jerd-jed | terch-tech | | | 2 | 4 | 8 |
| Same pairs | /iː/-/iː/ | theeb-theeb | teeve-teeve | feep-feep | beesh-beesh | 4 | 2 | 8 |
| | /ɪ/-/ɪ/ | thib-thib | tiv-tiv | fip-fip | bish-bish | 4 | 2 | 8 |
| | /æ/-/æ/ | jad-jad | gadge-gadge | jat-jat | dap-dap | 4 | 2 | 8 |
| | /ʌ/-/ʌ/ | jud-jud | gudge-gudge | jut-jut | dup-dup | 4 | 2 | 8 |
| | /ɜː/-/ɜː/ | derve-derve | jert-jert | jerd-jerd | terch-terch | 4 | 2 | 8 |
| | /ɑː/-/ɑː/ | darve-darve | jart-jart | | | 2 | 2 | 4 |
| | /ɛ/-/ɛ/ | jed-jed | tech-tech | | | 2 | 2 | 4 |
| | | | | | Total number of trials: 96 (48 same and 48 different) | | | |

*Note*: * Four possible combinations: two talker orders T1-T2, T2-T1, and two vowel orders: V1-V2, V2-V1.

## Appendix B Tukey pair-wise comparison results for each significant interaction.

**Table B1.** Identification results. Group by test interaction, group effect

| Tests of effect slices for Group*Test sliced by Test | | | | |
|---|---|---|---|---|
| Test | Num DF | Den DF | F Value | Pr > F |
| T1 Pretest | 2 | 23900 | 0.33 | 0.7188 |
| T2 Posttest | 2 | 23900 | 4.16 | 0.0155 |
| T3 Posttest2 | 0 | . | . | . |
| T4 Delayed test | 2 | 23900 | 3.42 | 0.0329 |

**Table B2.** Identification results. Group by test interaction, Tukey-corrected pairwise comparisons. Test comparisons within group. Coefficient estimates, standard errors, t-values, and adjusted *p*-values

| Simple effect level | Test | Test | Estimate | Standard error | t Value | Adj P |
|---|---|---|---|---|---|---|
| Group CG | T1 Pretest | T2 Posttest | –0.2492 | 0.1402 | –1.78 | 0.2844 |
| Group CG | T1 Pretest | T3 Posttest2 | –1.0585 | 0.1429 | –7.41 | <.0001 |
| Group CG | T1 Pretest | T4 Delayed test | –1.0035 | 0.1699 | –5.91 | <.0001 |
| Group CG | T2 Posttest | T3 Posttest2 | –0.8093 | 0.1432 | –5.65 | <.0001 |
| Group CG | T2 Posttest | T4 Delayed test | –0.7543 | 0.1702 | –4.43 | <.0001 |
| Group CG | T3 Posttest2 | T4 Delayed test | 0.05505 | 0.1725 | 0.32 | 0.9888 |
| Group DISG | T1 Pretest | T2 Posttest | –0.4784 | 0.1242 | –3.85 | 0.0003 |
| Group DISG | T1 Pretest | T4 Delayed test | –0.5750 | 0.1314 | –4.38 | <.0001 |
| Group DISG | T2 Posttest | T4 Delayed test | –0.09659 | 0.1319 | –0.73 | 0.7444 |
| Group IDG | T1 Pretest | T2 Posttest | –1.0707 | 0.1304 | –8.21 | <.0001 |
| Group IDG | T1 Pretest | T4 Delayed test | –1.1026 | 0.1440 | –7.66 | <.0001 |
| Group IDG | T2 Posttest | T4 Delayed test | –0.03189 | 0.1458 | –0.22 | 0.9740 |

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

**Table B3.** Identification results. Group by test interaction, Tukey-corrected pairwise comparisons. Group comparisons within test. Coefficient estimates, standard errors, t-values, and adjusted *p*-values

| Simple effect level | Group | Group | Estimate | Standard error | t Value | Adj P |
|---|---|---|---|---|---|---|
| Test T1 Pretest | CG | DISG | 0.2153 | 0.2687 | 0.80 | 0.7023 |
| Test T1 Pretest | CG | IDG | 0.1518 | 0.2732 | 0.56 | 0.8436 |
| Test T1 Pretest | DISG | IDG | –0.06349 | 0.2567 | –0.25 | 0.9668 |
| Test T2 Posttest | CG | DISG | –0.01394 | 0.2693 | –0.05 | 0.9985 |
| Test T2 Posttest | CG | IDG | –0.6697 | 0.2745 | –2.44 | 0.0390 |
| Test T2 Posttest | DISG | IDG | –0.6558 | 0.2582 | –2.54 | 0.0299 |
| Test T4 Delayed test | CG | DISG | 0.6437 | 0.2890 | 2.23 | 0.0666 |
| Test T4 Delayed test | CG | IDG | 0.05265 | 0.2970 | 0.18 | 0.9828 |
| Test T4 Delayed test | DISG | IDG | –0.5911 | 0.2688 | –2.20 | 0.0713 |

**Table B4.** Identification results. Word type by test interaction, effect of word type

| Tests of effect slices for Test*Word type sliced by Test | | | | |
|---|---|---|---|---|
| Test | Num DF | Den DF | F Value | Pr > F |
| T1 Pretest | 1 | 23900 | 1.32 | 0.2513 |
| T2 Posttest | 1 | 23900 | 3.95 | 0.0468 |
| T3 Posttest2 | 1 | 8174 | 2.08 | 0.1497 |
| T4 Delayed test | 1 | 23900 | 3.84 | 0.0500 |

**Table B5.** Identification results. Word type by test interaction, effect of test

| Tests of effect slices for Test*Word type sliced by Word type | | | | |
|---|---|---|---|---|
| Word type | Num DF | Den DF | F Value | Pr > F |
| nonword | 2 | 23900 | 36.45 | <.0001 |
| real word | 2 | 23900 | 63.40 | <.0001 |

**Table B6.** Identification results. Word type by test interaction, Tukey-corrected pairwise comparisons. Test comparisons within word type. Coefficient estimates, standard errors, t-values, and adjusted *p*-values

| Simple effect level | Test | Test | Estimate | Standard error | t Value | Adj P |
|---|---|---|---|---|---|---|
| Word type nonword | T1 Pretest | T2 Posttest | –0.4807 | 0.08380 | –5.74 | <.0001 |
| Word type nonword | T1 Pretest | T3 Posttest2 | –1.0066 | 0.1706 | –5.90 | <.0001 |
| Word type nonword | T1 Pretest | T4 Delayed test | –0.7762 | 0.09469 | –8.20 | <.0001 |
| Word type nonword | T2 Posttest | T3 Posttest2 | –0.8842 | 0.1707 | –5.18 | <.0001 |
| Word type nonword | T2 Posttest | T4 Delayed test | –0.2954 | 0.09511 | –3.11 | 0.0054 |
| Word type nonword | T3 Posttest2 | T4 Delayed test | 0.1598 | 0.2041 | 0.78 | 0.8622 |
| Word type real word | T1 Pretest | T2 Posttest | –0.7181 | 0.08559 | –8.39 | <.0001 |
| Word type real word | T1 Pretest | T3 Posttest2 | –1.1142 | 0.1745 | –6.38 | <.0001 |
| Word type real word | T1 Pretest | T4 Delayed test | –1.0112 | 0.09751 | –10.37 | <.0001 |
| Word type real word | T2 Posttest | T3 Posttest2 | –0.7368 | 0.1752 | –4.21 | 0.0002 |
| Word type real word | T2 Posttest | T4 Delayed test | –0.2931 | 0.09857 | –2.97 | 0.0083 |
| Word type real word | T3 Posttest2 | T4 Delayed test | –0.05712 | 0.2125 | –0.27 | 0.9932 |

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

**Table B7.** Discrimination results. Group by test interaction, group effect

| Tests of effect slices for Group*Test sliced by Test | | | | |
|---|---|---|---|---|
| Test | Num DF | Den DF | F Value | Pr > F |
| T1 Pretest | 2 | 22052 | 0.22 | 0.8022 |
| T2 Posttest | 2 | 22052 | 0.87 | 0.4192 |
| T3 Posttest2 | 0 | . | . | . |
| T4 Delayed test | 2 | 22052 | 0.49 | 0.6118 |

**Table B8.** Discrimination results. Group by test interaction, Tukey-corrected pairwise comparisons. Test comparisons within group. Coefficient estimates, standard errors, t-values, and adjusted *p*-values

| Simplee level | Test | Test | Estimate | Standard error | t Value | Adj P |
|---|---|---|---|---|---|---|
| Group CG | T1 Pretest | T2 Posttest | –0.1354 | 0.1096 | –1.23 | 0.6045 |
| Group CG | T1 Pretest | T3 Posttest2 | –0.4309 | 0.1118 | –3.85 | 0.0007 |
| Group CG | T1 Pretest | T4 Delayed Test | –0.5806 | 0.1353 | –4.29 | 0.0001 |
| Group CG | T2 Posttest | T3 Posttest2 | –0.2955 | 0.1127 | –2.62 | 0.0434 |
| Group CG | T2 Posttest | T4 DelayedTest | –0.4453 | 0.1363 | –3.27 | 0.0060 |
| Group CG | T3 Posttest2 | T4 DelayedTest | –0.1497 | 0.1380 | –1.09 | 0.6987 |
| Group DISG | T1 Pretest | T2 Posttest | –0.4672 | 0.09785 | –4.77 | <.0001 |
| Group DISG | T1 Pretest | T4 DelayedTest | –0.5993 | 0.1054 | –5.69 | <.0001 |
| Group DISG | T2 Posttest | T4 DelayedTest | –0.1321 | 0.1078 | –1.23 | 0.4382 |
| Group IDG | T1 Pretest | T2 Posttest | –0.5195 | 0.1018 | –5.10 | <.0001 |
| Group IDG | T1 Pretest | T4 DelayedTest | –0.4806 | 0.1114 | –4.31 | <.0001 |
| Group IDG | T2 Posttest | T4 DelayedTest | 0.03889 | 0.1141 | 0.34 | 0.9380 |

**Table B9.** Discrimination results. Group by test interaction, Tukey-corrected pairwise comparisons. Group comparisons within test. Coefficient estimates, standard errors, t-values, and adjusted *p*-values

| Simple effect level | Group | Group | Estimate | Standard error | t Value | Adj P |
|---|---|---|---|---|---|---|
| Test T1 Pretest | CG | DISG | 0.1170 | 0.2038 | 0.57 | 0.8341 |
| Test T1 Pretest | CG | IDG | 0.1234 | 0.2072 | 0.60 | 0.8226 |
| Test T1 Pretest | DISG | IDG | 0.006426 | 0.1944 | 0.03 | 0.9994 |
| Test T2 Posttest | CG | DISG | –0.2149 | 0.2057 | –1.04 | 0.5489 |
| Test T2 Posttest | CG | IDG | –0.2607 | 0.2094 | –1.25 | 0.4265 |
| Test T2 Posttest | DISG | IDG | –0.04583 | 0.1976 | –0.23 | 0.9708 |
| Test T4 DelayedTest | CG | DISG | 0.09833 | 0.2238 | 0.44 | 0.8991 |
| Test T4 DelayedTest | CG | IDG | 0.2235 | 0.2284 | 0.98 | 0.5907 |
| Test T4 DelayedTest | DISG | IDG | 0.1251 | 0.2066 | 0.61 | 0.8170 |

**Table B10.** Discrimination results. Word type by test interaction, effect of word type

| Tests of effect slices for Test*Word type sliced by Test | | | | |
|---|---|---|---|---|
| Test | Num DF | Den DF | F Value | Pr > F |
| T1 Pretest | 1 | 22052 | 1.96 | 0.1612 |
| T2 Posttest | 1 | 22052 | 5.47 | 0.0194 |
| T3 Posttest2 | 1 | 7542 | 5.78 | 0.0162 |
| T4 Delayed test | 1 | 22052 | 8.69 | 0.0032 |

https://doi.org/10.1017/S0272263124000408 Published online by Cambridge University Press

**Table B11.** Discrimination results. Word type by test interaction, effect of test

| Tests of effect slices for Test*Word type sliced by Word type | | | | |
|---|---|---|---|---|
| Word type | Num DF | Den DF | F Value | Pr > F |
| nonword | 2 | 22052 | 15.65 | <.0001 |
| real word | 2 | 22052 | 37.92 | <.0001 |

**Table B12.** Discrimination results. Word type by test interaction, Tukey-corrected pairwise comparisons. Test comparisons within word type. Coefficient estimates, standard errors, t-values, and adjusted p-values

| Simple effect level | Test | Test | Estimate | Standard error | t Value | Adj P |
|---|---|---|---|---|---|---|
| Word type nonword | T1 Pretest | T2 Posttest | –0.2923 | 0.07047 | –4.15 | <.0001 |
| Word type nonword | T1 Pretest | T3 Posttest2 | –0.4271 | 0.1289 | –3.31 | 0.0051 |
| Word type nonword | T1 Pretest | T4 Delayed Test | –0.4107 | 0.07954 | –5.16 | <.0001 |
| Word type nonword | T2 Posttest | T3 Posttest2 | –0.2944 | 0.1301 | –2.26 | 0.1068 |
| Word type nonword | T2 Posttest | T4 DelayedTest | –0.1184 | 0.08075 | –1.47 | 0.3071 |
| Word type nonword | T3 Posttest2 | T4 DelayedTest | –0.1076 | 0.1576 | –0.68 | 0.9038 |
| Word type real word | T1 Pretest | T2 Posttest | –0.4558 | 0.07414 | –6.15 | <.0001 |
| Word type real word | T1 Pretest | T3 Posttest2 | –0.4306 | 0.1395 | –3.09 | 0.0110 |
| Word type real word | T1 Pretest | T4 DelayedTest | –0.6963 | 0.08578 | –8.12 | <.0001 |
| Word type real word | T2 Posttest | T3 Posttest2 | –0.2944 | 0.1412 | –2.08 | 0.1582 |
| Word type real word | T2 Posttest | T4 DelayedTest | –0.2405 | 0.08805 | –2.73 | 0.0173 |
| Word type real word | T3 Posttest2 | T4 DelayedTest | –0.1882 | 0.1755 | –1.07 | 0.7061 |

**Cite this article:** Cebrian, J., Gavaldà, N., Gorba, C., & Carlet, A. (2024). Differential effects of identification and discrimination training tasks on L2 vowel identification and discrimination. *Studies in Second Language Acquisition,* 46: 1069–1093. https://doi.org/10.1017/S0272263124000408